BI169: COMPUTATIONAL METHODS IN MOLECULAR & CELLULAR BIOLOGY
Winter/Spring 1998; California Institute of Technology, Pasadena, CA
(http://www.bbb.caltech.edu/bowerlab/Bi169/)

GENETIC NETWORK INFERENCE IN COMPUTATIONAL MODELS AND APPLICATIONS TO LARGE-SCALE GENE EXPRESSION DATA

Roland Somogyi, Ph.D

Molecular Physiology of CNS Development
LNP/NINDS/NIH, 36/2C02, Bethesda, MD 20892
P: 301-402-1407
e-mail: rolands@helix.nih.gov
W^3: http://rsb.info.nih.gov/mol-physiol/homepage.html

Extended Abstract

Large scale gene expression mapping is motivated by the premise that the information on the functional state of an organism is largely determined by the information on gene expression (based on the central dogma). This process may be conceptualized as a genetic feedback network, in which information flows from gene activity patterns through a hierarchy of inter- and intracellular signaling functions back to the regulation of gene expression. The genetic network perspective is particularly relevant to the study of development. Gene sequence information in cis regions (regulatory inputs) and protein coding regions (regulatory outputs; determines biomolecular dynamics) is expanded into spatio-temporal structures defining the organism. Some essentials of this behavior may be captured by computational models, such as Boolean networks.

In order to draw meaningful inferences from gene expression data, it is important that each gene is surveyed under several different conditions, preferably time series. Such data sets may be analyzed using a range of methods with increasing depth of inference, such as cluster analysis, and determination of mutual information content. Abstract computational models may serve as a test bed for the development of these inference techniques. Only in such models can the dynamic behavior of many elements (trajectories, attractors) be unequivocally linked to a selected network architecture (wiring and rules). Beyond cluster analysis, complete genetic network reverse engineering has been established for such idealized models.

We have applied some of the more rudimentary inference techniques to an extensive expression survey of selected gene families in CNS development. Detailed cluster analysis has uncovered typical waves of expression, characterizing distinct phases of development of spinal cord and hippocampus. Indeed, distinct functional classes and gene families clearly map to particular expression profiles, suggesting that the definition of pathways may be recast in terms of gene expression clusters. These pathways may be likened to modules in genetic programs. Analysis of gene activity patterns following injury-induced responses in hippocampus (kainic acid induced seizures and excitotoxic cell death) indicates a general "recapitulation of developmental programs". Comprehensive reconstruction of genetic networks should be further facilitated by integration of gene expression data with knowledge of cis-regulatory structures and other criteria of gene function, once available.

Section I

"Learn to walk with the model before you run with the data."

Note: The model network trajectories and basins of attraction shown below were generated using the DDLAB software by Andy Wuensche, Ph.D.

- Andy Wuensche's web site: http://www.santafe.edu/~wuensch/

Complementary perspectives

A Yin perspective:

"A model of a system is only as good as the predictions that it makes that would not have been seen otherwise." - Mel Simon

The Yang perspective:

However, new, intriguing, but speculative models lead to the search and discovery of new data.

A Yang perspective:

Experimental data on a system is only as good as the conceptual framework or model within which we can interpret its meaning.

The Yin perspective:

However, new and surprising experimental data leads to the development of models that revolutionize our thinking.

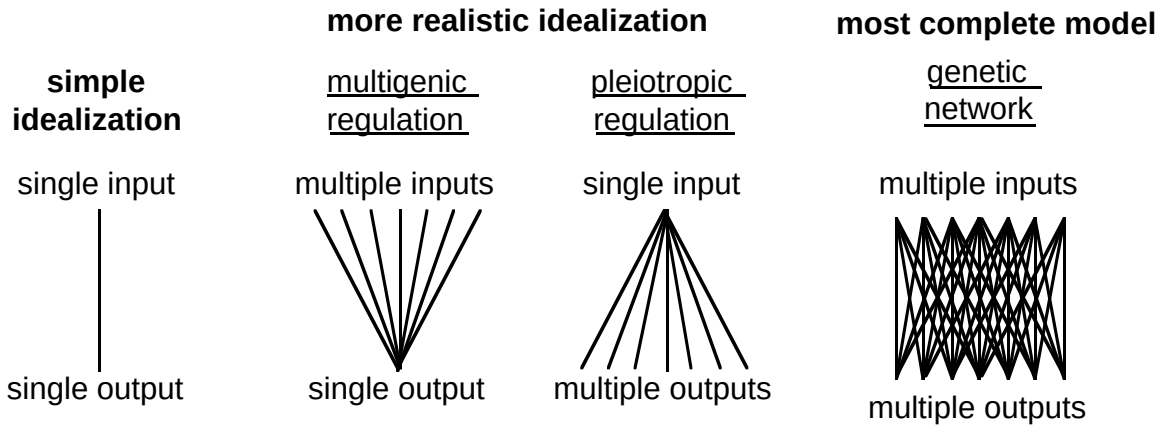
They key uncertainties:

When do we decide that a speculative model is worth investigating?

When do we decide that an unusual experiment is worth conducting?

...usually, in both cases, after the fact ;-)

Multigenic & pleiotropic regulation: the basis of genetic networks

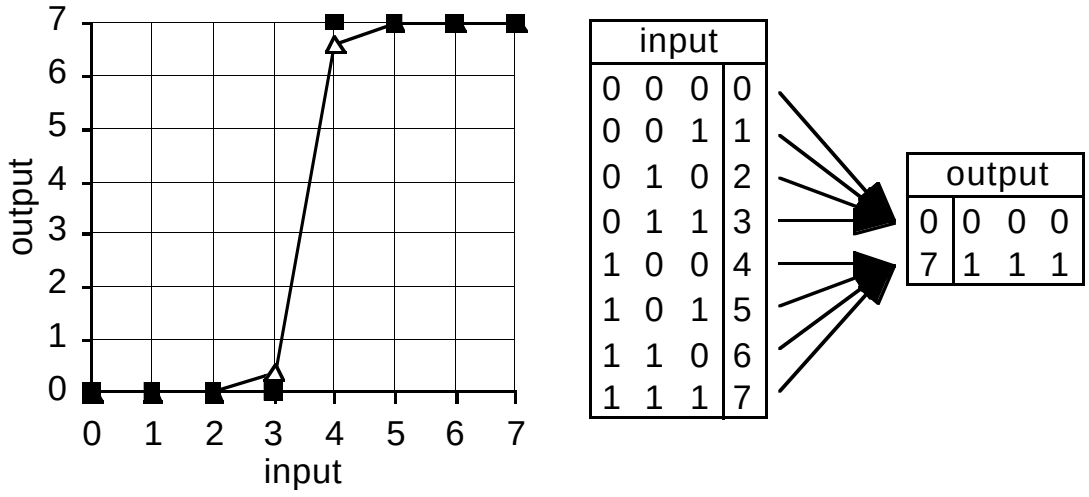


Somogyi R, Fuhrman S, Askenazi M, Wuensche A (1997) The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. Nonlinear Analysis, Proc. of Second World Cong. of Nonlinear Analysts (WCNA96), 30(3):1815-1824.

How can we conceptualize a distributed biomolecular network?

Perhaps a model based on idealized, elemental mechanisms can illustrate the nature of complex behavior:

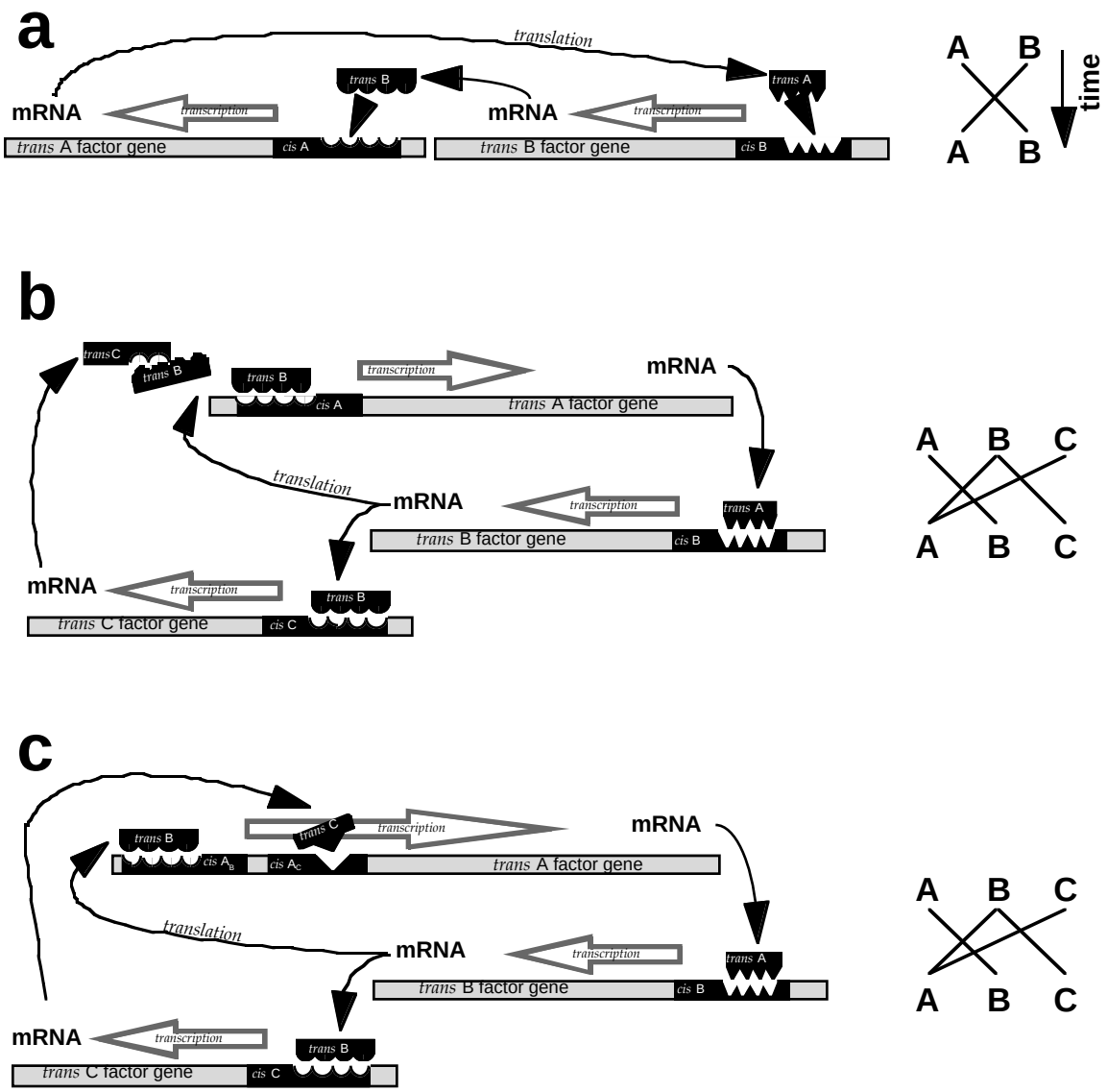
- Each gene may receive one or several inputs from other genes or itself.
- Assuming a highly cooperative, sigmoid input-output relationship, a gene can be modeled as a binary element.



Somogyi, R (1998) Many to One Mappings as a Basis for Life.Interjournal (in press)

- The output (time=t+1) is computed from the input (time=t) according to logical or Boolean rules. Time is discrete, and all genes are updated simultaneously.

Cis / trans interactions constitute the basis of network wiring

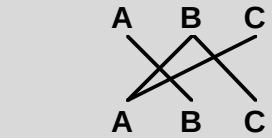


Somogyi & Sniegoski,1996; Complexity 1(6):45-63

Genes code for trans-acting proteins, which in turn control the expression of genes through interactions with cis regulatory sites located on the DNA molecule. The cybernetic foundations of such networks are represented by wiring diagrams, shown on the right, and the computational rules determining the input/output relationships. a) Positive feedback system between genes A and B. b) and c) Same as a), except that C inhibits and overrides the stimulatory action of B on A. This is accomplished by protein-protein (b) or protein-DNA (c) interactions, which are computationally equivalent.

Wiring and rules determine network dynamics

Wiring and rules



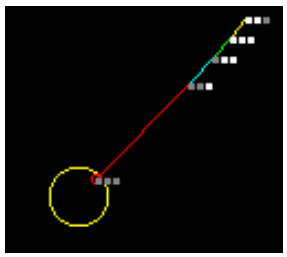
inputs	2	1	1
rule	4	2	2

Basis for rules:

1. A activates B
2. B activates A and C
3. C inhibits A

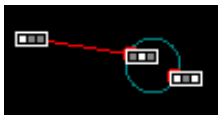
Trajectory 1 results in a point attractor

iteration	A	B	C
1	1	1	0
2	1	1	1
3	0	1	1
4	0	0	1
5	0	0	0
6	0	0	0



Trajectory 2 results in a 2-state dynamic attractor

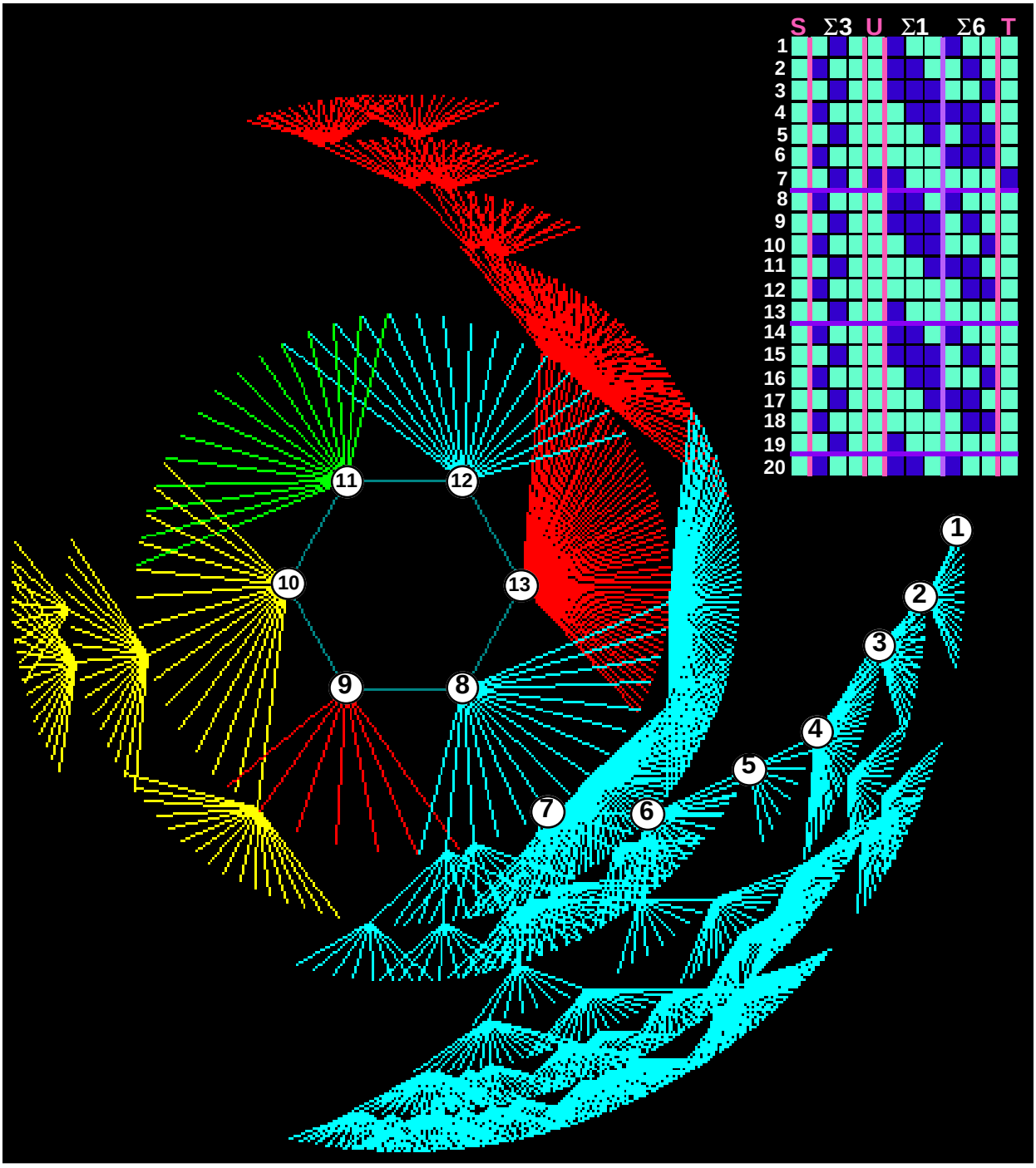
iteration	A	B	C
1	1	0	0
2	0	1	0
3	1	0	1
4	0	1	0



Somogyi & Sniegoski,1996; Complexity 1(6):45-63

Wiring diagram (top panel): The lines connect the upper row of output elements (time=t) to the lower row of input elements (time=t+1). The no. of inputs and the pertaining decimal rule are shown underneath each wiring diagram. Time space patterns or trajectories (lower panels) can be directly calculated from the wiring and rules. Middle panel: point attractor, basin includes 5 states. Lower panel: 2-state dynamic attractor (repeating pattern), basin includes 3 states. The basins of attraction include all 8 possible states of the system.

Many states converge on one attractor



Somogyi & Sniegoski,1996; Complexity 1(6):45-63

The network trajectory (upper right panel) inexorably leads to a final state or state cycle - an attractor (center). Each state of the trajectory is shown as a point (labeled by its time step number). The labeled trajectory (state points connected by lines) is one of many trajectories leading to the repeating, six-state attractor pattern. The centripetal trajectories leading to the attractor form the basin of attraction. Minor state perturbations will not change the final outcome of the network, conferring stability.

Network terminology

	Architecture	
wiring	<->	biomolecular connections
rules (functions, codes)	<->	biomolecular interactions
	Dynamics	
state	<->	set of molecular activity values; e.g. gene expression, signaling molecules
state transition	<->	response to previous state
trajectory	<->	series of state transitions; e.g. differentiation, perturbation response
attractor	<->	final outcome; e.g. phenotype, cell type, chronic illness

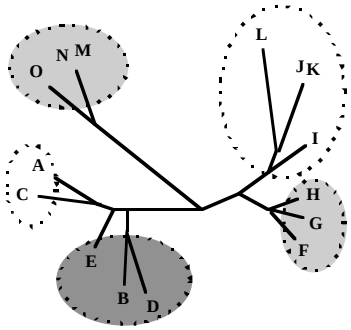
Inference of shared control processes:

- Similarities in gene expression patterns suggest shared control.
- Clustering gene expression patterns according to a heuristic distance measure is the first step toward constructing a wiring diagram.
- Euclidean distance as a measure for the difference between gene expression patterns: A gene expression pattern over n time points is a point in n-dimensional parameter space.

  Distance = $\sqrt{\Sigma (a_i - b_i)^2}$

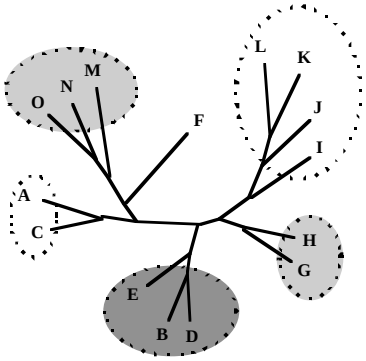
Euclidean cluster analysis of a model network

Wiring (Molecular Interaction) Clusters



gene	Boolean rule
A	F and H and J
B	G and H and J
C	F and H and I
D	G and H and I
E	H and I and J
F	I and J and K and L and (not G)
G	I and J and K and L and (not O)
H	I and J and K and L
I	J and K and L
J	K and L
K	K or L
L	L or M
M	N or O
N	N and O
O	N and O and (not E)

Trajectory (Gene Expression) Clusters



trajectory	I											II				III				IV			
time	1	2	3	4	5	6	7	8	9	10	1	2	3	4	1	2	3	4	1	2	3	4	
A	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0	0	0	
B	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0	
C	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0	1	0	0	0	
D	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	0	0	1	1	0	0	
E	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	
F	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	
G	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0	
H	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0	
I	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0	
J	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	
K	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
L	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
M	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
O	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Somogyi R, Fuhrman S, Askenazi M, Wuensche A (1997) The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. Nonlinear Analysis, Proc. of Second World Cong. of Nonlinear Analysts (WCNA96), 30(3):1815-1824.

Note that the clustering pattern in the functional time series (lower panel) closely resembles the gene groupings according to wiring (upper panel).

Biological information flow

"It's not about mechanism, it's about information flow."

- But, what is information?
- Can information be quantified?
- Can information measures be used in network analysis?

Information can be quantified: Shannon entropy (H)

$H(X) = -\Sigma\, p_x \log p_x$ $H(Y) = -\Sigma\, p_y \log p_y$ $H(X,Y) = -\Sigma\, p_{x,y} \log p_{x,y}$

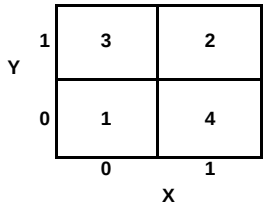
a

X	0	1	1	1	1	1	1	0	0	0
Y	0	0	0	1	1	0	0	1	1	1

H(X) = -0.4log(0.4)-0.6log(0.6) = 0.97 *(40% 0s and 60% 1s)*

H(Y) = -0.5log(0.5)-0.5log(0.5) = 1.00 *(50% 0s and 50% 1s)*

b

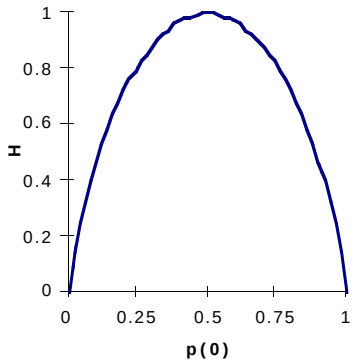


H(X,Y) = -0.1log(0.1)-0.4log(0.4)-0.3log(0.3)-0.2log(0.2) = 1.85

Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

Determination of H. a) Single element. Probabilities (p) are calculated from frequency of on/off values of X and Y. b) Distribution of value pairs. H is calculated from the probability of co-occurrence of x, y values over all measurements.

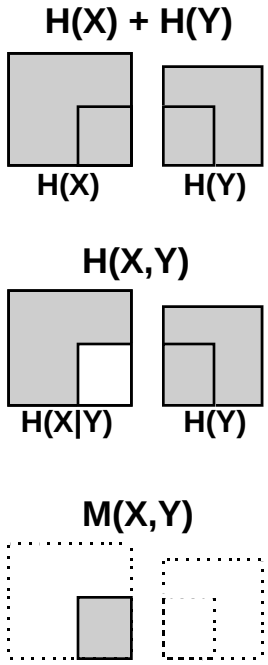
The Shannon entropy is maximal if all states are equiprobable



Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

Shannon entropies for a 2-state information source. Since the sum of the state probabilities must be unity, p(1)=1-p(0) for 2 states.

Mutual information (M):
the information (Shannon entropy) shared by non-independent elements

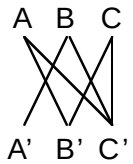


Liang S, Fuhrman S, Somogyi R (1998) Pacific Symposium on Biocomputing 3:18-29.

Venn diagrams of information relationships. In each case, add the shaded portions of both squares to determine one of the following: [H(X)+H(Y)], H(X,Y), and M(X,Y). The small corner rectangles represent information that X and Y have in common. H(Y) is shown smaller than H(X) and with the corner rectangle on the left instead of the right to indicate that X and Y are different, although they have some mutual information.

A candidate Boolean network for reverse engineering

a



b

A' = B
B' = A or C
C' = (A and B) or (B and C) or (A and C)

c

input			output		
A	**B**	**C**	**A'**	**B'**	**C'**
0	0	0	0	0	0
0	0	1	0	1	0
0	1	0	1	0	0
0	1	1	1	1	1
1	0	0	0	1	0
1	0	1	0	1	1
1	1	0	1	1	1
1	1	1	1	1	1

Liang S, Fuhrman S, Somogyi R (1998) Pacific Symposium on Biocomputing 3:18-29.

a) Wiring diagram. b) Boolean rules. c) State transition table; input column shows all states at time=t, outputs (prime) correspond to the matching states at time=t+1.

The principle behind REVEAL

(REVerse Engineering ALgorithm)

Input entropies

H(A)	1.00
H(B)	1.00
H(C)	1.00
H(A,B)	2.00
H(B,C)	2.00
H(A,C)	2.00
H(A,B,C)	3.00

$$H(X) = -\sum p(x) \log p(x)$$
$$H(X,Y) = -\sum p(x,y) \log p(x,y)$$

$$M(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$M(X,[Y,Z]) = H(X) + H(Y,Z) - H(X,Y,Z)$$

Determination of inputs for element A ①

H(A')	1.00				
H(A',A)	2.00	M(A',A)	0.00	M(A',A) / H(A')	0.00
H(A',B)	1.00	M(A',B)	1.00	**M(A',B) / H(A')**	**1.00**
H(A',C)	2.00	M(A',C)	0.00	M(A',C) / H(A')	0.00

② **Rule table for A**

rule no. 2

input	output
B	**A'**
0	0
1	1

Determination of inputs for element B ③

H(B')	0.81				
H(B',A)	1.50	M(B',A)	0.31	M(B',A) / H(B')	0.38
H(B',B)	1.81	M(B',B)	0.00	M(B',B) / H(B')	0.00
H(B',C)	1.50	M(B',C)	0.31	M(B',C) / H(B')	0.38
H(B',[A,B])	2.50	M(B',[A,B])	0.31	M(B',[A,B]) / H(B')	0.38
H(B',[B,C])	2.50	M(B',[B,C])	0.31	M(B',[B,C]) / H(B')	0.38
H(B',[A,C])	2.00	M(B',[A,C])	0.81	**M(B',[A,C]) / H(B')**	**1.00**

④ **Rule table for B**

rule no. 14

input		output
A	**C**	**B'**
0	0	0
0	1	1
1	0	1
1	1	1

Determination of inputs for element C ⑤

H(C')	1.00				
H(C',A)	1.81	M(C',A)	0.19	M(C',A) / H(C')	0.19
H(C',B)	1.81	M(C',B)	0.19	M(C',B) / H(C')	0.19
H(C',C)	1.81	M(C',C)	0.19	M(C',C) / H(C')	0.19
H(C',[A,B])	2.50	M(C',[A,B])	0.50	M(C',[A,B]) / H(C')	0.50
H(C',[B,C])	2.50	M(C',[B,C])	0.50	M(C',[B,C]) / H(C')	0.50
H(C',[A,C])	2.50	M(C',[A,C])	0.50	M(C',[A,C]) / H(C')	0.50
H(C',[A,B,C])	3.00	M(C',[A,B,C])	1.00	**M(C',[A,B,C]) / H(C')**	**1.00**

⑥ **Rule table for C**

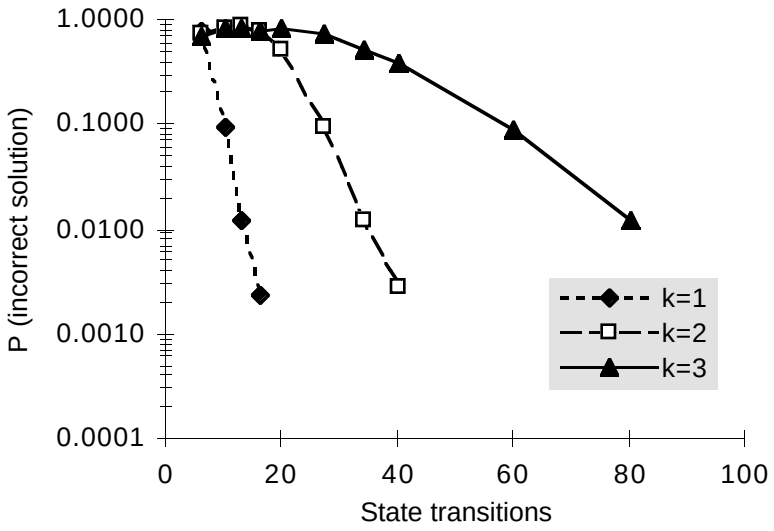
rule no. 170

input			output
A	**B**	**C**	**C'**
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Liang S, Fuhrman S, Somogyi R (1998) Pacific Symposium on Biocomputing 3:18-29.

Hs and Ms are calculated from the time series or look-up tables according to the definitions (shaded). The wiring of the example Boolean network can be inferred from the state transition table using progressive M-analysis (left, odd steps). Once the inputs (wiring) to a gene are know, one can construct the rule table by matching the states of the inputs to those of the output from the state transition table (right, even steps).

Inference from incomplete time series or state transition tables



Liang S, Fuhrman S, Somogyi R (1998) Pacific Symposium on Biocomputing 3:18-29.

REVEAL will quickly find a minimal solution for a Boolean network given any set of time series. For n=50 (genes) and k=3 or less (number of inputs per gene), the correct or full solution can be unequivocally inferred from 100 state transition pairs. Note that for n=50 (genes) and k=3 (inputs per gene) only a small fraction (~100) of all possible state transitions ($2^{50} \sim 10^{15}$!) is required for reliable inference of the network wiring and rules.

Suggested reading for SECTION I

Somogyi R, Sniegoski CA (1996) Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. Complexity 1(6):45-63.

Somogyi R, Fuhrman S, Askenazi M, Wuensche A (1997) The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. Nonlinear Analysis, Proc. of Second World Cong. of Nonlinear Analysts (WCNA96), 30(3):1815-1824.

Somogyi, R (1998) Many to One Mappings as a Basis for Life. Interjournal (in press).

Liang S, Fuhrman S, Somogyi R (1998) REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

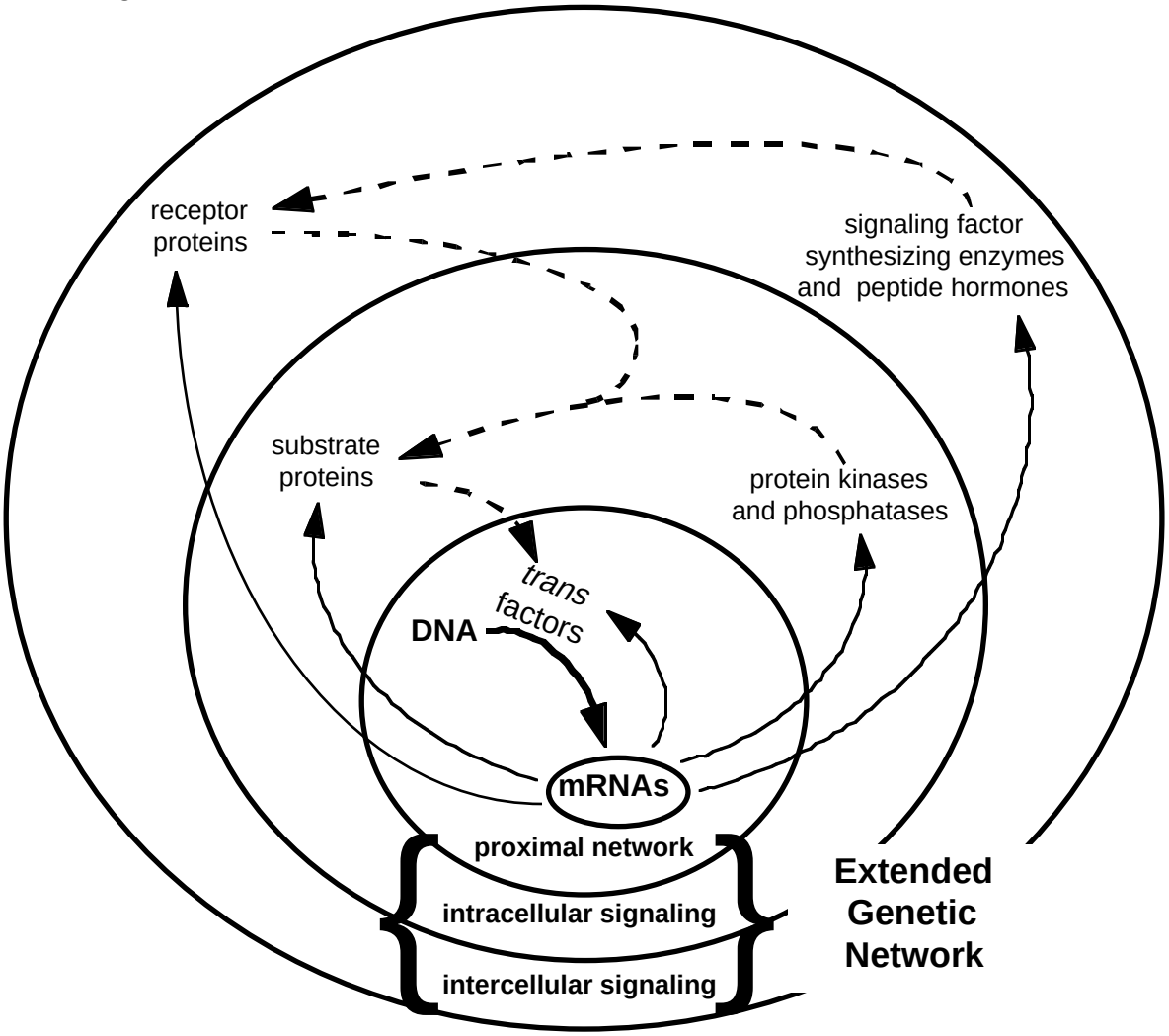
Somogyi R., Fuhrman, S (1997) Distributivity, a General Information Theoretic Network Measure, or Why the Whole is More than the Sum of its Parts. Proceedings of the International Workshop on Information Processing in Cells and Tissues 1997 (in press).

Section II

"You're on your own: start running (i.e. with the large-scale expression data)"

Information flow in genetic networks

- Genes regulate the expression of genes through a hierarchy of signaling functions.
- Gene expression patterns represent the variables, while the signaling functions are determined by the gene structure.
- In this feedback network, the expression of mRNA can be viewed as both, the origin and target of informaiton flow.



Somogyi & Sniegoski,1996; Complexity 1(6):45-63

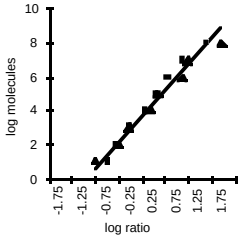
The solid lines refer to information flow from primary sources (DNA, mRNA). The broken lines correspond to information flow from secondary sources back to the primary source (Somogyi & Sniegoski,1996; *Complexity* 1(6):45-63).

Functional inference from large-scale gene expression data

- Gene expression patterns contain much of the state information of the system and can be measured experimentally.
- We are facing the challenge of reverse engineering the internal structure of this genetic signaling network from measurements of its output.
- This will require high precision in data acquisition, and sufficient coherence among the data sets, as found in time series.

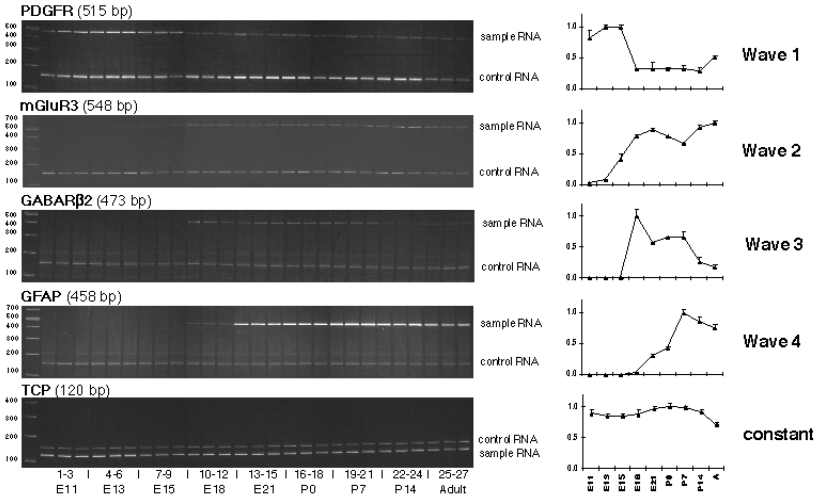
High precision, high sensitivity assay

- RT-PCR (reverse transcription polymerase chain reaction)
- Flexible and scalable through automation
- RNA standard serves as internal control
- Measurement scales linearly with RNA copy number on log scales



Somogyi R, Wen X, Ma W, Barker JL (1995) Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord. J Neurosci 15:2575-2591

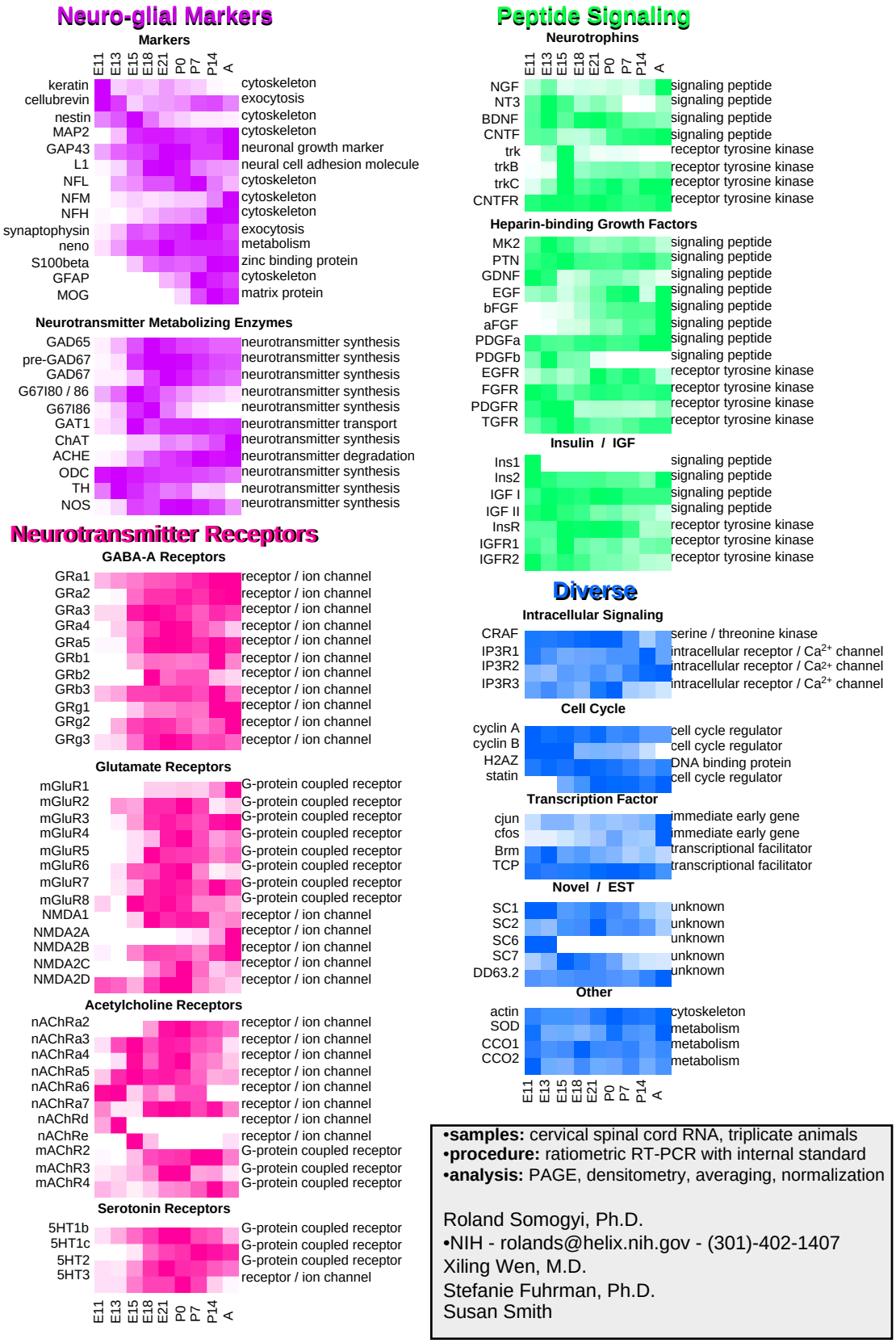
RT-PCR analysis of gene expression in developing rat CNS



Roland Somogyi, Ph.D. Stefanie Fuhrman, Ph.D.
Xiling Wen, M.D. Susan Smith

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998) Large-Scale Temporal Gene Expression Mapping of CNS Development. Proc Natl Acad Sci USA, 95:334-339.

The Gene Expression Matrix of rat spinal cord development (PNAS 95:334-339)



- **samples:** cervical spinal cord RNA, triplicate animals
- **procedure:** ratiometric RT-PCR with internal standard
- **analysis:** PAGE, densitometry, averaging, normalization

Roland Somogyi, Ph.D.
•NIH - rolands@helix.nih.gov - (301)-402-1407
Xiling Wen, M.D.
Stefanie Fuhrman, Ph.D.
Susan Smith

Inference of shared control processes

Cluster analysis

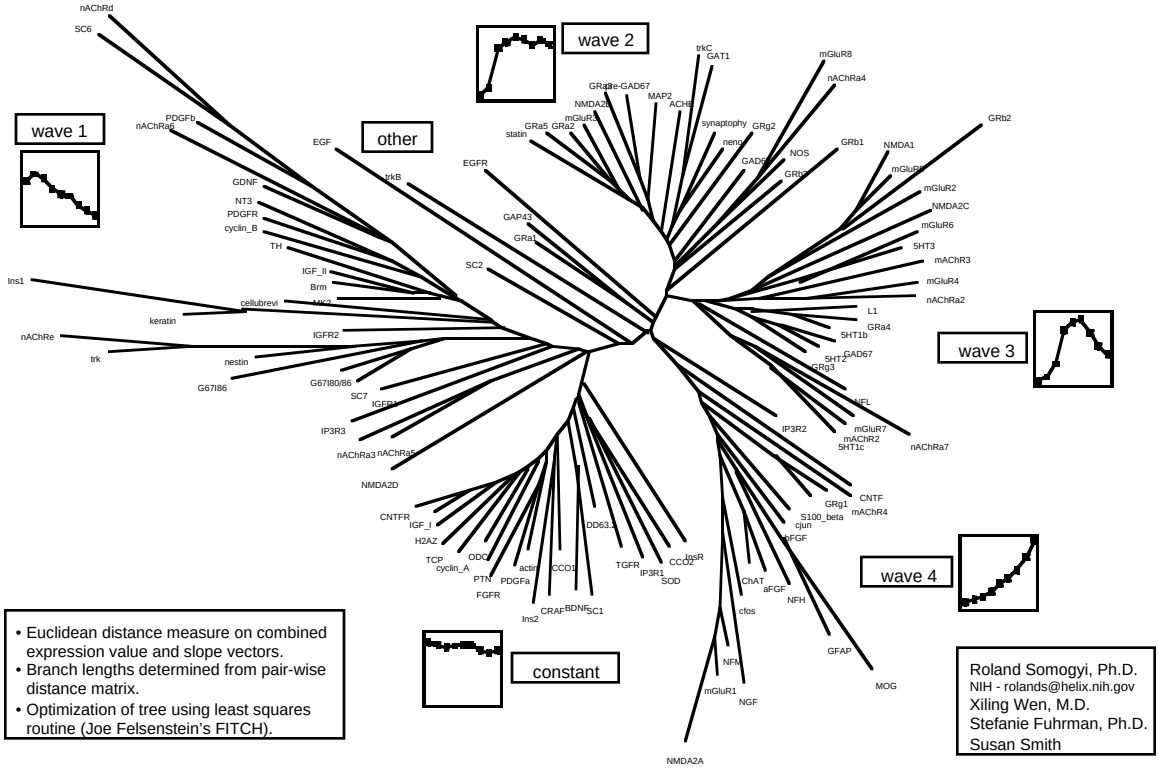
- Similarities in gene expression patterns suggest shared control.
- Clustering gene expression patterns according to a heuristic distance measure is the first step toward constructing a wiring diagram.

Complete reverse engineering

- Only possible for model networks with current algorithms and limited data sets.
- Careful experimental designs and optimizaiton of inference tools may allow significant progress in the future.

Euclidean cluster analysis of gene expression time series in spinal cord development

- Euclidean distance: A gene expression pattern over n time points is a point in n-dimensional parameter space:
  Distance = $\sqrt{\Sigma (a_i - b_i)^2}$
- Euclidean cluster analysis implies shared wiring and rules.



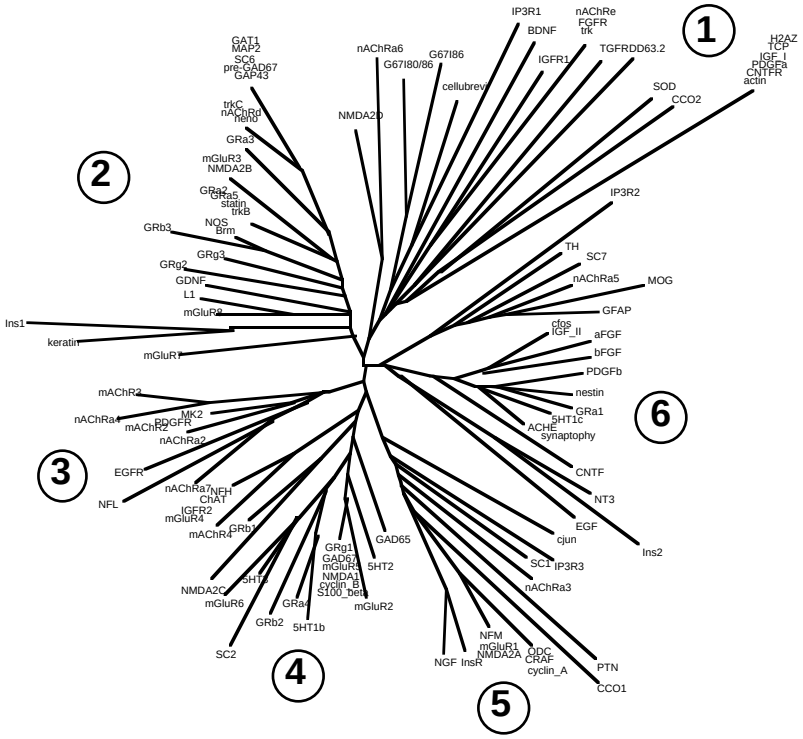
- Euclidean distance measure on combined expression value and slope vectors.
- Branch lengths determined from pair-wise distance matrix.
- Optimization of tree using least squares routine (Joe Felsenstein's FITCH).

Roland Somogyi, Ph.D.
NIH - rolands@helix.nih.gov
Xiling Wen, M.D.
Stefanie Fuhrman, Ph.D.
Susan Smith

Carr DB, Somogyi R, Michaels G (1997) Templates for Looking at Gene Expression Clustering. Statistical Computing and Graphics Newsletter 8(1):20-29.

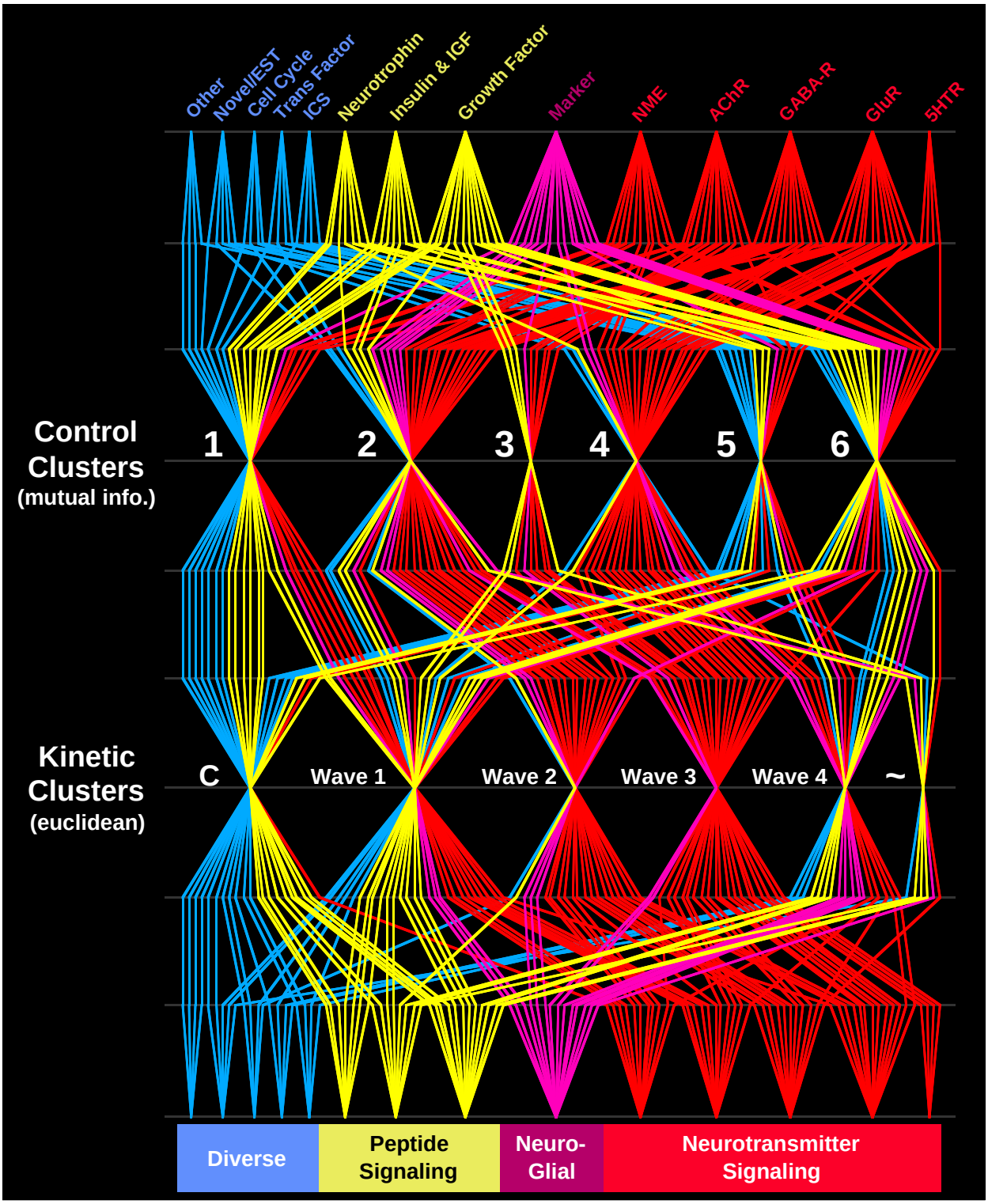
Mutual information cluster analysis

- Mutual information: Most general measure of correlation:
  M = H(A) + H(B) - H(A, B)
- "Coherence" (normalized mutual information): Captures similarities in patterns independent of individual information entropies. "In how far is pattern A able to predict pattern B?":
  C = M(A, B) / Hmax(A, B)
- **Mutual information (Coherence) cluster analysis implies shared wiring, with no constraints on rules.**



Michaels G, Carr DB, Wen X, Fuhrman S, Askenazi M, Somogyi R (1998) Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. Pacific Symposium on Biocomputing 3:42-53.

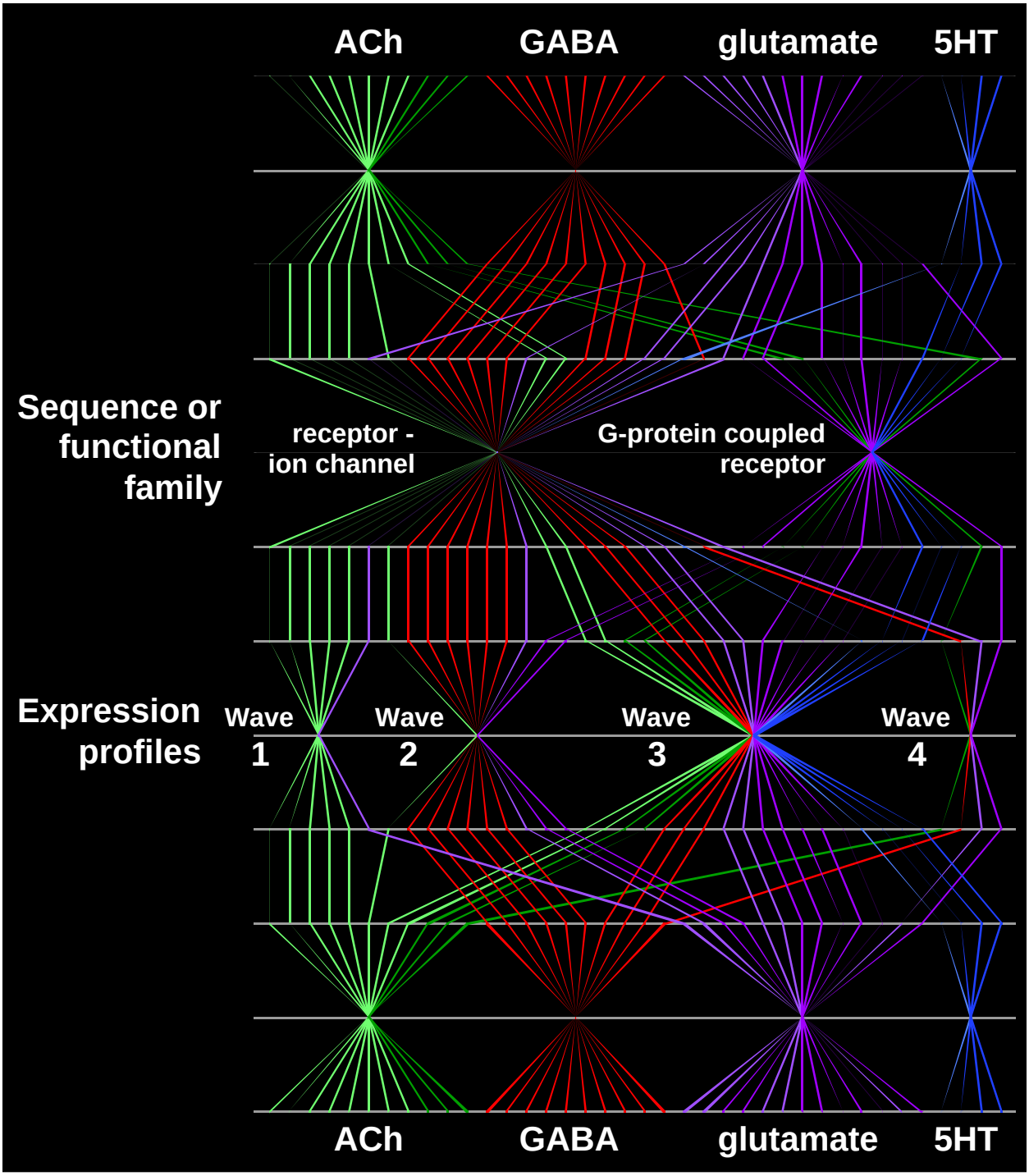
Functional gene families map to distinct control processes



Michaels G, Carr DB, Wen X, Fuhrman S, Askenazi M, Somogyi R (1998) Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. Pacific Symposium on Biocomputing 3:42-53.

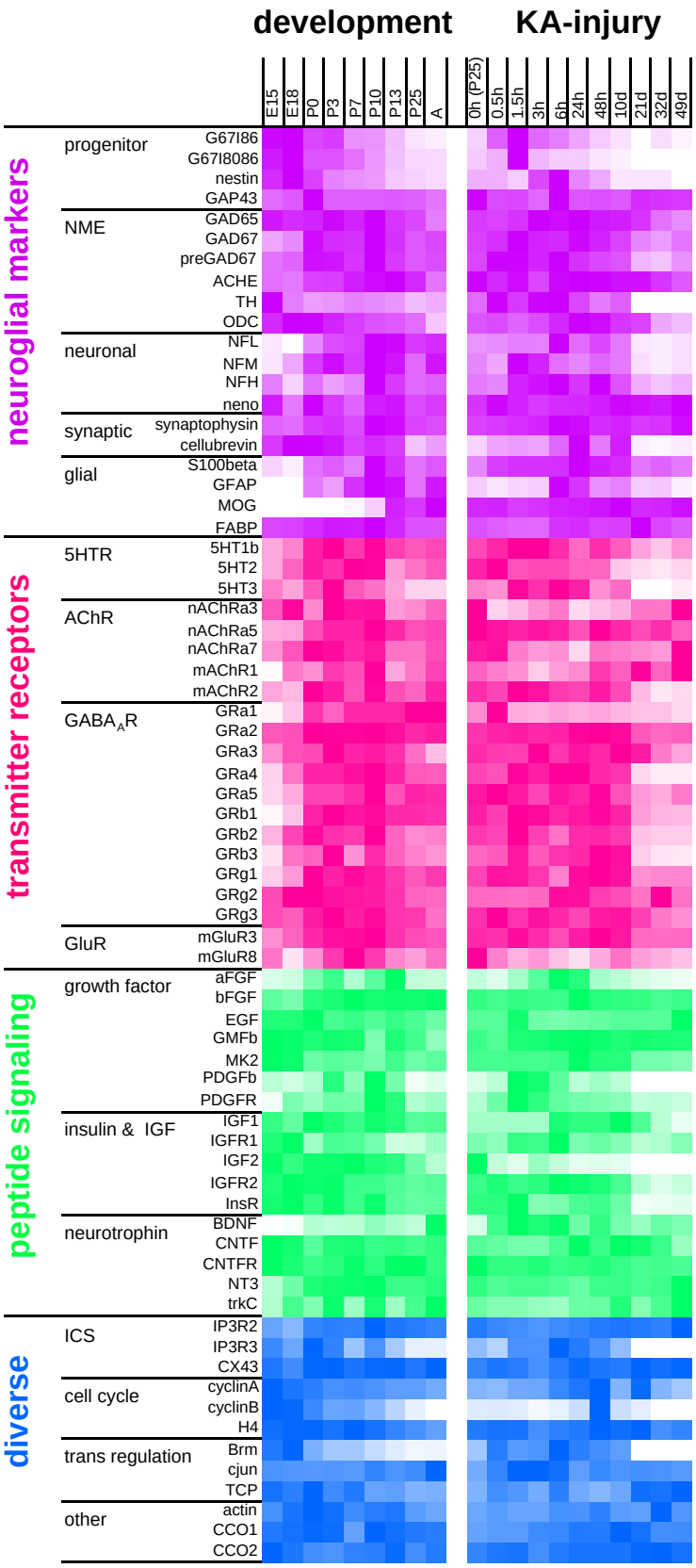
Note that the functional gene families map to specific expression pathways, e.g. the peptide signaling genes (yellow) do not appear in waves 2 and 3.

Neutrotransmitter receptors follow particular expression waveforms according to ligand and functional class.

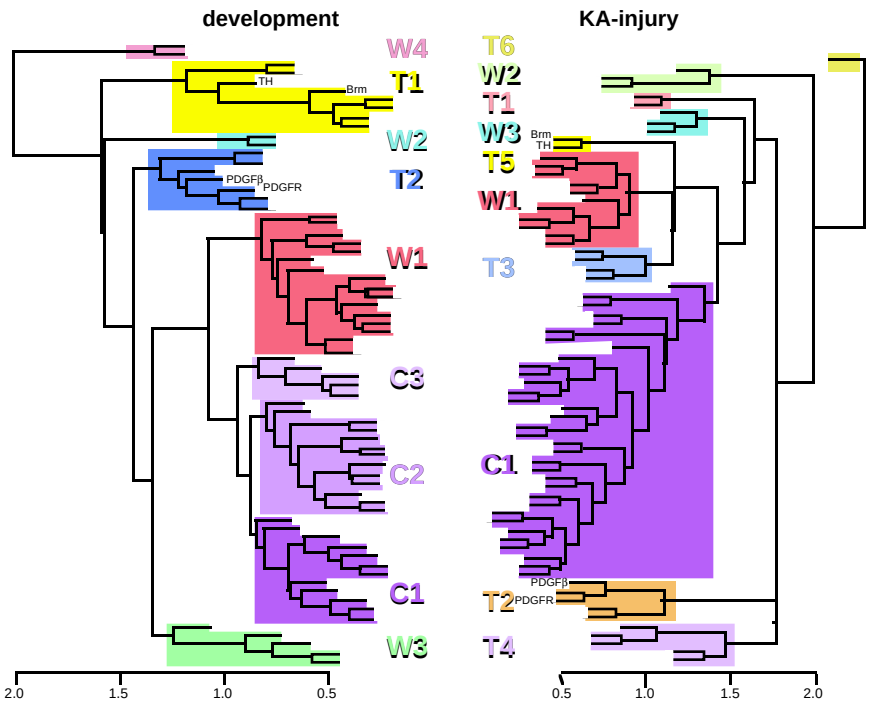


See euclidean distance tree (above) for pictograms showing typical expression profile for each wave. Note that the early expression waves 1 and 2 are dominated by Ach and GABA receptors, and by receptor ion-channels in general.

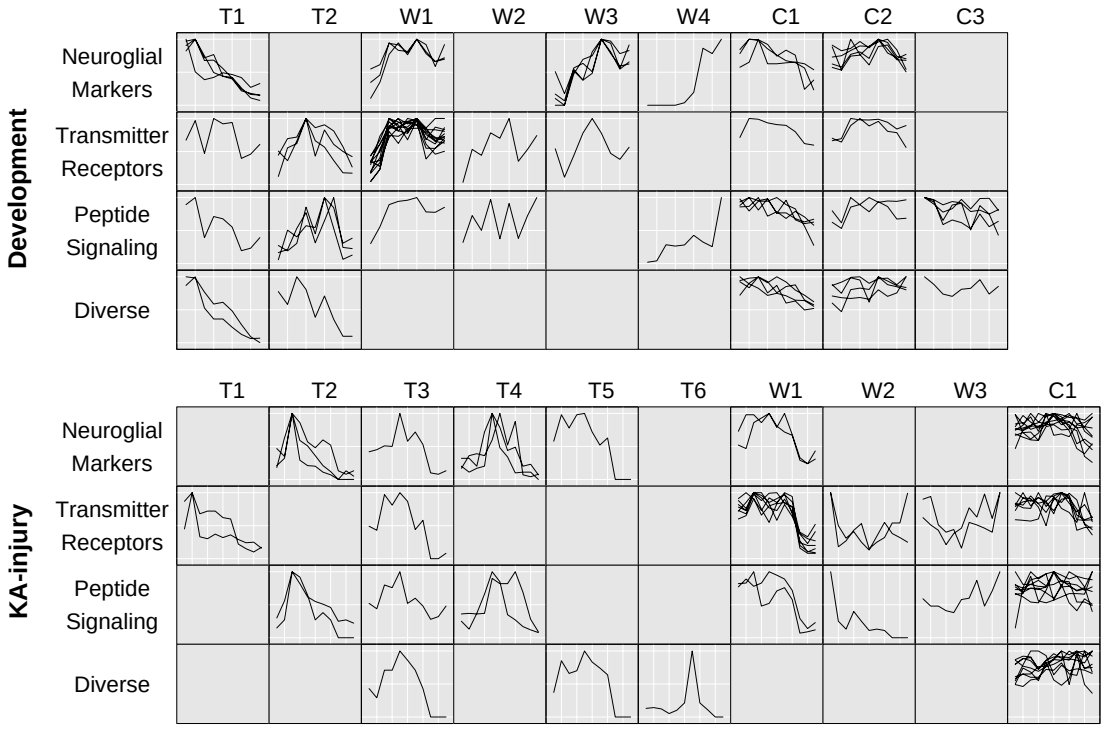
Gene expression matrix of hippocampal development and injury.



Cluster analysis of hippocampal gene expression time series.

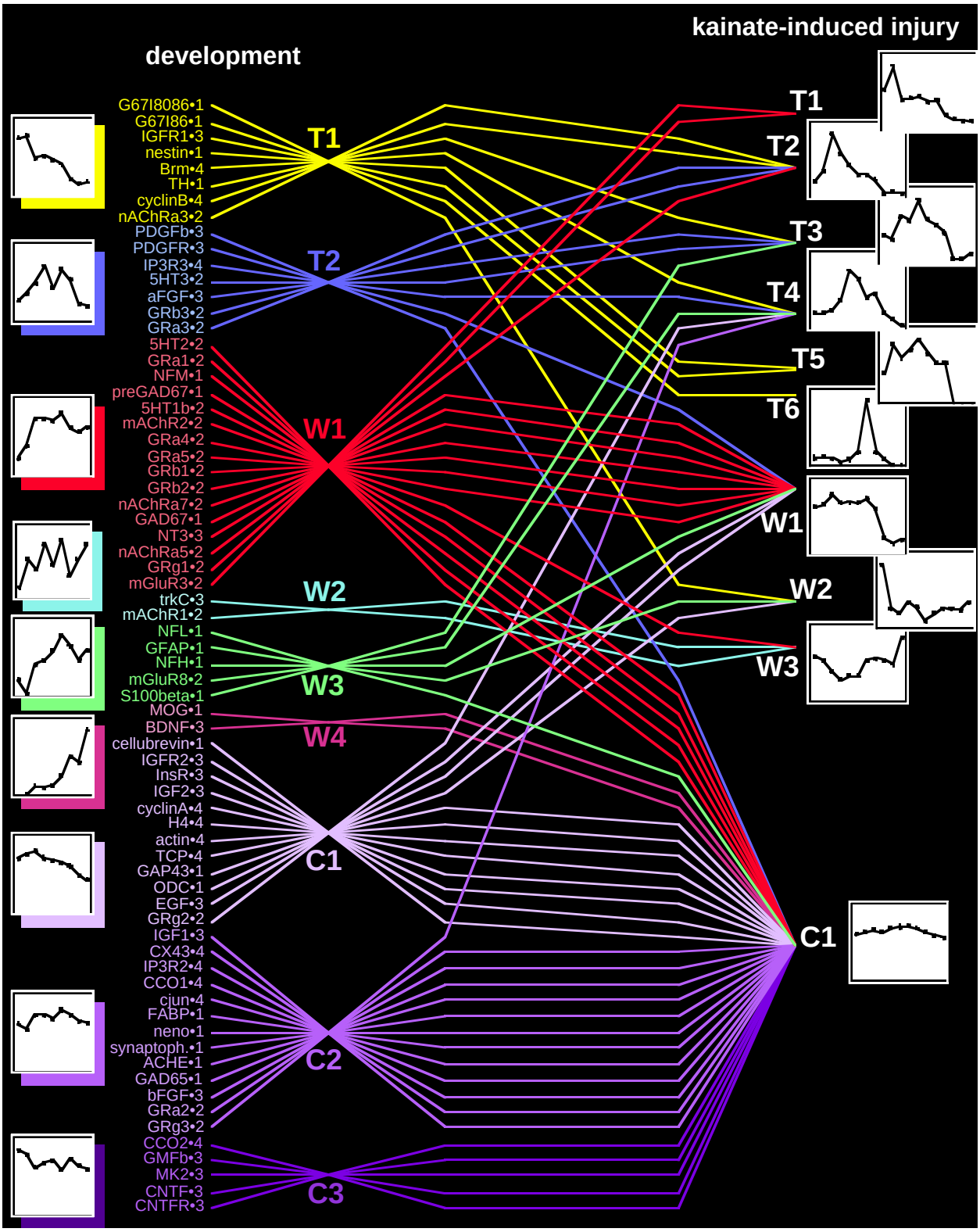


Distribution of gene expression profiles – analysis of cluster tightness



Mapping of developmental gene expression clusters to KA-injury clusters

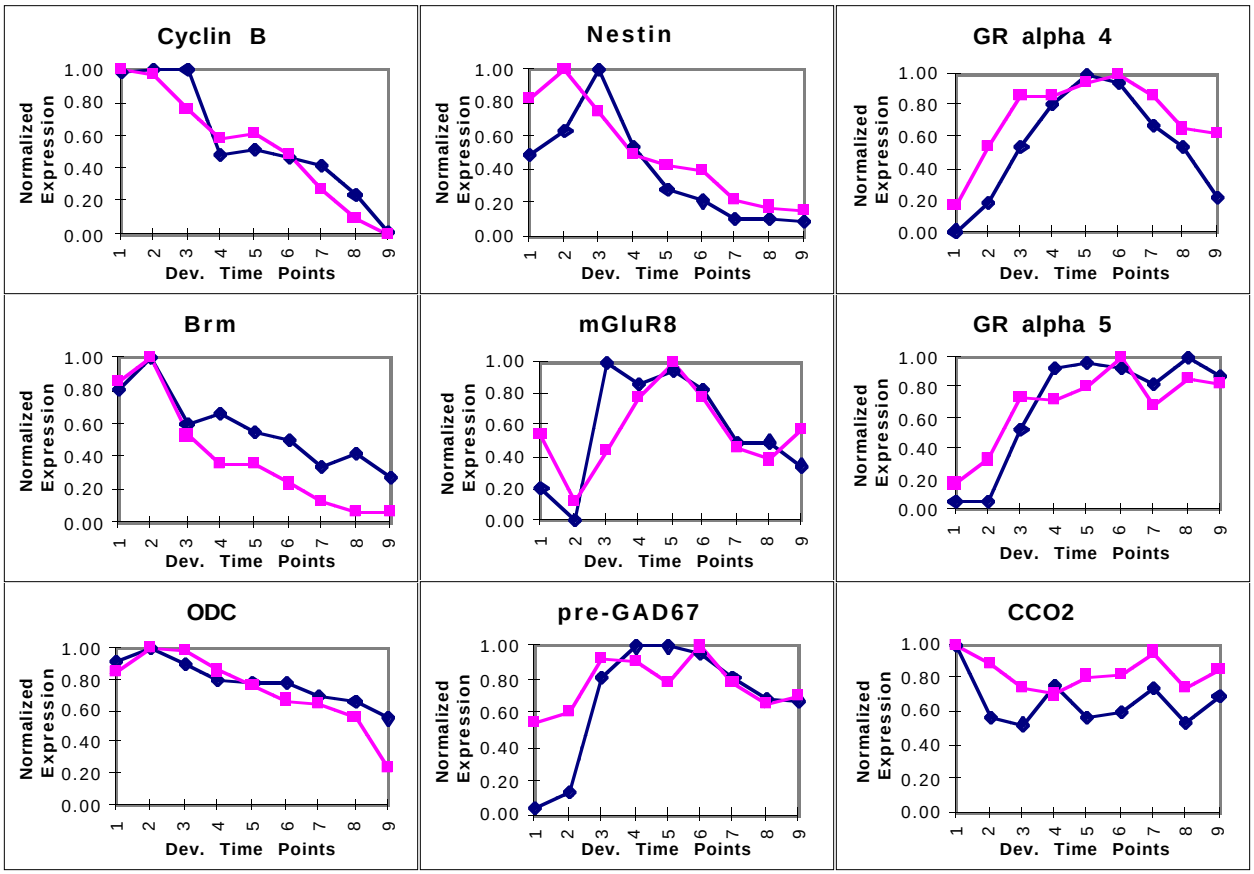
"Recapitulation of developmental programs"



Caption of previous graph:

Average expression patterns for all clusters are shown as pictograms. Colors correspond to developmental clusters (matches shadowing of developmental pictograms). Lines connect genes in developmental clusters to their respective KA-injury clusters. Each gene can be followed from its label (left column) along a line connecting it to the first focus (developmental cluster) and then, according to the mirror image of this line, to the focus of the KA-injury cluster. Clusters are labeled by Ts, Ws and Cs, corresponding to "Transient", "Waveform" and "Constant" patterns. In development, Ts mark genes that are expressed significantly higher during early to mid development in relation to adult, Ws characterize genes that show other fluctuating patterns, and Cs mark clusters that are relatively high in expression over all time points. During response to KA-injury, Ts correspond to genes that temporarily increase in expression, Ws to genes that fluctuate according to alternative patterns, and Cs to genes that remain relatively high in expression over all time points. Note that T, W and C cluster members in development generally map to the corresponding T, W and C patterns following KA-injury.

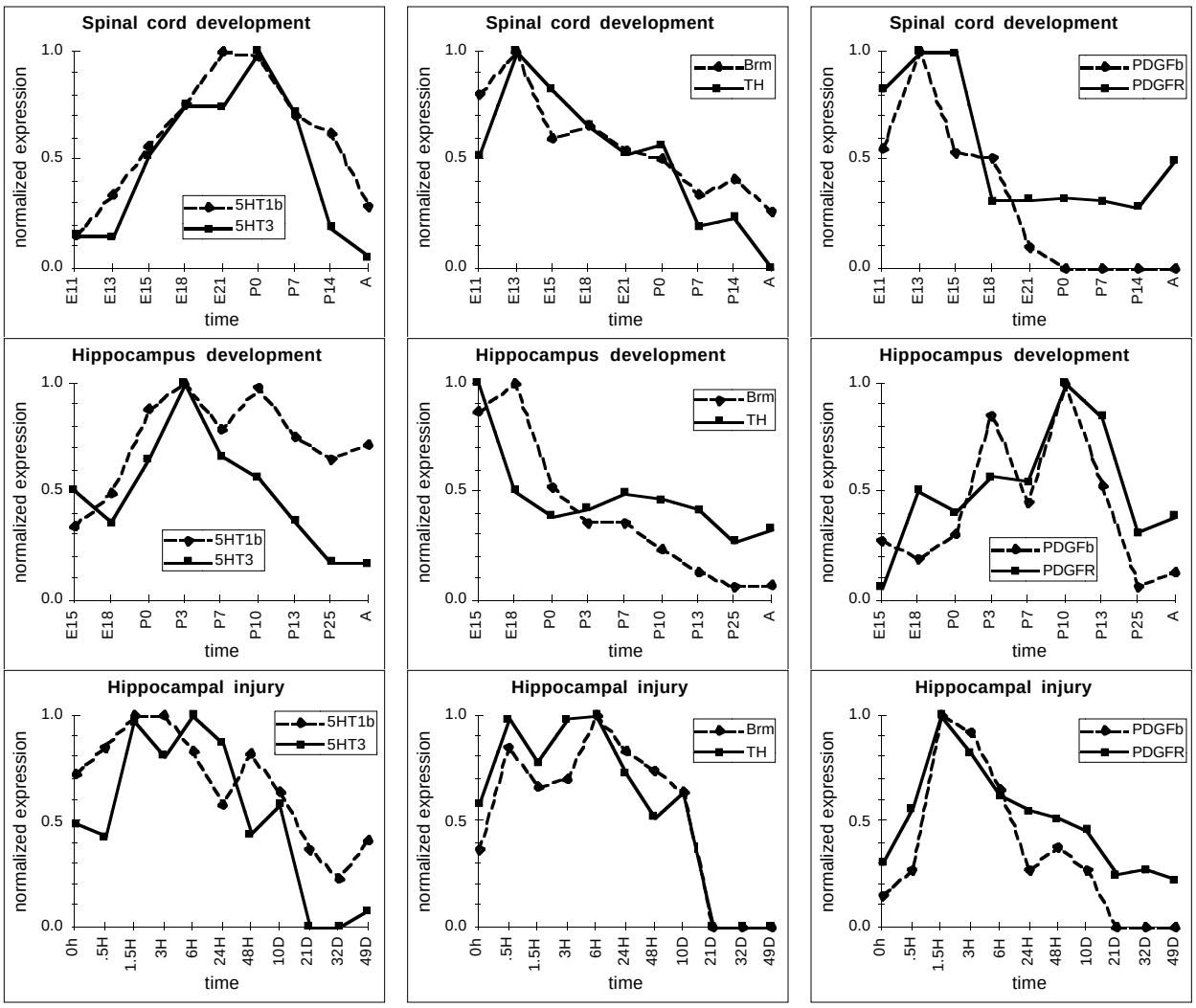
Overlapping control of gene expression in spinal cord and hippocampus



(blue=spinal cord expression; pink=hippocampal expression)

Analysis of CNS development and injury identifies tightly co-regulated genes: evidence for genetic programs

5HT1β R	(metabotropic)	Brm	(transcription)	PDGF β	(peptide)
5HT3 R	(ionotropic)	TH	(enzyme)	PDGF R	(receptor)



same ligand, different family
similar control in s.c. and hippo.

no known gene relationship
similar control in s.c. and hippo.

peptide / receptor pair
unique control in s.c. and hippo.

- The similarity of gene expression patterns between spinal cord and hippocampus development and hippocampal injury suggests the existence **generalized genetic programs** common to all CNS regions.
- The assumption that this finding can be extrapolated to other CNS structures is not far-fetched given the evolutionary distance between hippocampus, a structure derived from cerebral cortex, and spinal cord.
- The reactivation of developmental genes after injury implies a general recapitulation of developmental programs

Suggested reading for SECTION II

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998) Large-Scale Temporal Gene Expression Mapping of CNS Development. Proc Natl Acad Sci USA, 95:334-339.

Michaels G, Carr DB, Wen X, Fuhrman S, Askenazi M, Somogyi R (1998) Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. Pacific Symposium on Biocomputing 3:42-53.

Carr DB, Somogyi R, Michaels G (1997) Templates for Looking at Gene Expression Clustering. Statistical Computing and Graphics Newsletter 8(1):20-29.

D'haeseleer P, Wen X, Fuhrman S, Somogyi R (1997) Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data. Proceedings of the International Workshop on Information Processing in Cells and Tissues 1997 (in press).

Summary

General strategies for network model construction

Bottom up approach

- Determine characteristics of individual biomolecular interactions.
- Build model and test under new experimental conditions.

Top down approach

- Determine input-output patterns (time series) of network.
- Infer connections and rules using level-by-level inference.

Hybrid approach

- Knowledge of individual biomolecular interactions can serve as constraints that will accelerate reverse engineering

Level-by-level inference from large-scale gene expression data

Data requirements

- High precision measurement method
- Data must resemble time series or state transitions

Inference of shared control processes or pathways

- Euclidean distance analysis: shared wiring and rules
- Mutual information analysis: shared wiring, varying rules

Complete reverse engineering

- Established for simple logical networks
- The principle of REVEAL could be applied to experimental data
- Accurate analysis requires detailed measurements of extended time series in response to various perturbations
- Depth of inference is dependent on volume and resolution of data

What are we learning from theoretical studies?

- Principles of network architecture and dynamics.
- Distributed networks exhibit stability and robustness.
- Rule constraints greatly influence stability.
- Data analysis & inference techniques:
  Cluster analysis captures shared wiring and rules.
  Complete reverse engineering is possible for model networks. The limitations of this approach lie in the simplifying assumptions of the network, not its size!

What are we learning from experimental data?

- Biological networks are far from random in terms of their wiring and rules.
- There is a great deal of coherence of gene expression patterns within higher organizational structures involving multiple cell types.
- Constraints are being placed on network architecture from each level of biological organization.
- Genetic programs exhibit a distributed modular structure.
- Many genetic programs appear to be "variations on a theme" of a root program.

Outlook

The limits of Boolean networks as genetic network models

Simultaneous updating of all inputs

- It is not realistic in a biological network that all elements cross their thresholds synchronously. The updating order will affect the trajectory of the network (See work by Glass, Thieffry, Thomas).
- Solution: The updating order in a discrete network can be determined by the modeler. This has been implemented in the latest version of the DDLAB software. Alternatively, models using piecemeal differential equations have been implemented (Leon Glass).

Binary states

- Thresholding in biological systems is usually short of the idealized on/off behavior. Moreover, different processes may involve varying thresholds.
- Solution: The number of discrete states could be increased to cover every threshold that might be relevant. This would not change the nature of the discrete network, but would make it larger and more complex.

Missing elements

- This genetic networks model just focuses on gene expression. Of course, information is also carried through protein expression, phorsphorylation cascades, signaling molecules etc.
- Solution: This is not really a limit of the model, but rather due to lack of biological knowledge of all these parameters. However, due the informational redundancy, not all biological parameters may be required to provide a detailed outline of the network.

Comparison of alternative modeling frameworks

- Binary discrete network (simple, can handle large numbers of elements, oversimplification)
- Multi-state discrete network (approaches behavior of continuous network given sufficient state resolution; tradeoff in simplicity, more realistic)
- Continuous network (systems of differential equations, difficult to implement for large numbers of elements).
  Has been used succesfully to model the genetic network output responsible for patterning in the early fly embryo (Reinitz & collaborators) and lambda phage decision network (McAdams & collaborators).
- Exhaustive bottom-up modeling or top-down inference of biomolecular networks will probably be impossible. However, due to limited independence between different informational compartments in the organism, encapsulated aspects of the network may be amenable to predictive modeling.

Integration of top-down with bottom-up approaches

- Databases cataloging individual functional gene interactions
- Databases cataloging expression data (developmental time series, perturbations, tissues, cell types)
- Gene sequence analyis in terms of cis and ORF structural relationships
- Paradigms for integrating functional knowledge with top-down inference approaches
- Definition and determination of "root" programs

Why is this important?

It's the inevitable next step

- Progress in molecular biology, physiology and biochemistry clearly points to highly cross-wired networks. The one gene, one function, one connection perspective is plainly inadequate.
- Integration of large-scale biology with computational technology

Vital practical applications depend on integration

- Diagnosis and therapy of complex diseases
- Cancer
- Regeneration after injury
- Degenerative disorders

Engineering organisms

- Agriculture (growth, resistance, metabolic engineering)
- Microorganisms (waste treatment, chemical engineering)

Additional reading for SUMMARY

Overview

- Kauffman, S.A. (1993) The Origins of Order, Self-Organization and Selection in Evolution. Oxford University Press
- Bryant, A. Milosavljevic and R. Somogyi (1998) Gene Expression and Genetic Networks. Pacific Symposium on Biocomputing 3:3-5 (1998).

Asynchronous Boolean networks

- Thieffry, D. and Thomas, R. (1998) Qualitative Analysis of Gene Networks. Pacific Symposium on Biocomputing 3:66-76.
- Thomas, R. (1991) Regulatory Networks Seen as Asynchronous Automata: A Logical Description. J Theor Biol. 153: 1-23.

Continuous logical networks

- Glass, L. and Kauffman, S.A. (1973) The Logical Analysis of Continuous, Non-Linear Biochemical Control Networks. J. Theor. Biol. 39:103-129.
- Glass, L. (1975). Classification of Biological Networks by Their Qualitative Dynamics. J. Theor. Biol. 54:85-107.

Reverse engineering

- Arkin, A., Shen, P., Ross, J. (1997) A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. Science 277:1275-1279.
- Arkin, A., and Ross, J. (1997) Statistical Construction of Chemical Reaction Mechanisms from Measured Time-Series. J. Phys. Chem. 99:970-979.

Bottom-up modeling of small genetic networks

- McAdams, H.H., Shapiro, S. (1995) Circuit Simulation of Genetic Networks. Science. 269:650-656.

Stochastic behavior of networks

- McAdams, H.H., Arkin, A. (1997) Stochastic Mechanisms in Gene Expression. PNAS, USA 94(3):814.

Genetic network modeling of spatio-temporal patterns in development (combined bottom up and top-down approach)

- Mjolsness, E., Sharp, D. H. and Reinitz, J. (1991) A connectionist model of development. J. Theor. Biol. 152: 429-453.
- Reinitz, J., Mjolsness, E., and Sharp, D.H. (1995) Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. J. Exp. Zool. 271: 47-56.
- Reinitz, J. and Sharp, D.H. (1995) Mechanism of eve stripe formation. Mech. Dev.49: 133-158.

Cis-regulatory structures

- Arnone, M.I. and Davidson, E. (1997) The Hardwiring of Development: Organization and Function of Genomic Regulatory Systems. Development 124:1851-1864.

Biological constraints on genetic feedback networks

- Savageau, M.A. (1998) Rules for the Evolution of Gene Circuitry. Pacific Symposium on Biocomputing 3:54-65.

Project Participants

Staff

Xiling Wen - Large scale RTPCR analysis of gene expression patterns

Stefanie Fuhrman - Modeling, experimental design, data analysis

Susan Smith - PCR product analysis

Collaboratoring Scientists

George Michaels - Bioinformatics, George Mason University, Virginia.

Daniel Carr - George Mason University, Virginia

Shoudan Liang - NASA Ames Research Center, California

Patrick d'Haeseleer - Department of Computer Science, University of New Mexico

Millicent Dugich-Djordjevic - Laboratory of Developmental Neurobiology, NICHD, NIH

Manor Askenazi - Santa Fe Institute